

# On the non-ergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming

Ya-Feng Liu · Xin Liu · Shiqian Ma

Received: date / Accepted: date

**Abstract** In this paper, we consider the linearly constrained composite convex optimization problem, whose objective is a sum of a smooth function and a possibly nonsmooth function. We propose an inexact augmented Lagrangian (IAL) framework for solving the problem. The proposed IAL framework requires solving the augmented Lagrangian (AL) subproblem at each iteration less accurately than most of the existing IAL frameworks/methods. We analyze the global convergence and the non-ergodic convergence rate of the proposed IAL framework.

**Keywords** Inexact augmented Lagrangian framework · Non-ergodic convergence rate

**Mathematics Subject Classification (2000)** 90C25 · 65K05

## 1 Introduction

In this paper, we consider the linearly constrained composite convex optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & F(x) := f(x) + g(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{1.1}$$

where  $A \in \mathbb{R}^{m \times n}$  and  $b \in \mathbb{R}^m$ ;  $f(x)$  is a convex smooth function with Lipschitz continuous gradient; and  $g(x)$  is a closed convex (not necessarily smooth) function

---

Y.-F. Liu and X. Liu  
LSEC, ICMSEC, Academy of Mathematics and Systems Science,  
Chinese Academy of Sciences, Beijing 100190, China  
E-mail: yaffiu@lsec.cc.ac.cn; liuxin@lsec.cc.ac.cn

S. Ma  
Department of Systems Engineering and Engineering Management,  
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China  
E-mail: sqma@se.cuhk.edu.hk

with a bounded domain<sup>1</sup>. An important example of problem (1.1) is that  $g(x)$  is an indicator function of a convex compact set  $\mathcal{X}$ , that is,

$$g(x) = \text{Ind}_{\mathcal{X}}(x) := \begin{cases} 0, & \text{if } x \in \mathcal{X}, \\ +\infty, & \text{otherwise.} \end{cases}$$

In this case, problem (1.1) can be rewritten as

$$\min_{x \in \mathcal{X}} f(x), \text{ s.t. } Ax = b.$$

One efficient approach to solving problem (1.1) is the AL method [6, 19, 20]. The AL function of problem (1.1) is

$$\mathcal{L}_{\beta}(x; \lambda) := \hat{f}_{\beta}(x; \lambda) + g(x), \quad (1.2)$$

where

$$\hat{f}_{\beta}(x; \lambda) := f(x) + \langle \lambda, Ax - b \rangle + \frac{\beta}{2} \|Ax - b\|^2,$$

$\lambda \in \mathbb{R}^m$  is the Lagrange multiplier associated with the linear constraint, and  $\beta > 0$  is the penalty parameter. The (augmented) Lagrangian dual of problem (1.1) is

$$\max_{\lambda \in \mathbb{R}^m} d(\lambda) \quad (1.3)$$

with

$$d(\lambda) := \min_{x \in \mathbb{R}^n} \mathcal{L}_{\beta}(x; \lambda). \quad (1.4)$$

It is well-known that the dual function  $d(\lambda)$  in (1.4) is differentiable and its gradient is given by  $\nabla d(\lambda) = Ax(\lambda) - b$ , where  $x(\lambda)$  is the solution of problem (1.4) (see [1]).

Given  $\lambda^k$ , the AL method for solving problem (1.1) updates the primal and dual variables via

$$x(\lambda^k) = \arg \min_{x \in \mathbb{R}^n} \mathcal{L}_{\beta}(x; \lambda^k) \quad (1.5)$$

and

$$\lambda^{k+1} = \lambda^k + \beta \left( Ax(\lambda^k) - b \right),$$

respectively. The AL method for solving problem (1.1) is essentially a dual gradient ascent method, which updates the dual variable by performing a dual gradient ascent step

$$\lambda^{k+1} = \lambda^k + \beta \nabla d(\lambda^k).$$

When the problem dimension  $n$  is large, finding an exact solution of AL subproblem (1.5) can be computationally expensive and thus the exact gradient  $\nabla d(\lambda^k)$  is often unavailable. As a result, many works focused on inexact versions

---

<sup>1</sup> The bounded domain assumption on  $g(x)$  is made in this paper mainly for ease of presentation. For problem (1.1) arising from many applications of interest such as machine learning, statistics, and signal processing, we often can easily find a bounded set  $\mathcal{X}$  such that the solution of problem (1.1) lies in  $\mathcal{X}$ . Therefore, we can restrict the definition of  $g(x)$  over this bounded set. Let us take the compressed sensing problem, which is a special case of problem (1.1) with  $f(x) = 0$  and  $g(x) = \|x\|_1$ , as an example. We can restrict the definition of  $\|x\|_1$  over the bounded domain  $\{x \mid \|x\|_1 \leq \|\hat{x}\|_1\}$ , where  $\hat{x}$  is any point satisfying  $A\hat{x} = b$ .

of (dual) gradient methods; see [2–4, 10–14, 20–22] and references therein. Rockafellar [20] proposed an IAL framework, where the AL subproblem is solved until a point  $x^{k+1}$  is found such that

$$\mathcal{L}_\beta(x^{k+1}; \lambda^k) - \mathcal{L}_\beta(x(\lambda^k); \lambda^k) \leq \eta_k^2, \quad (1.6)$$

and showed that the proposed IAL framework converges if the nonnegative tolerance sequence  $\{\eta_k\}$  is summable. Very recently, Devolder, Glineur, and Nesterov [3] proposed a general inexact gradient framework and analyzed the ergodic convergence rate of their framework when it is applied to solve dual problem (1.3). In [14], Nedelcu, Necoara, and Tran-Dinh proposed an IAL method, where the AL subproblem was approximately solved by Nesterov’s gradient method [15–17] such that (1.6) is satisfied and showed again the ergodic convergence rate of the proposed IAL method. The non-ergodic convergence rate result for the IAL framework/method has been missing in the literature for a long time until in a very recent work by Lan and Monterio [10], where they proposed an IAL method (where the AL subproblems are approximately solved by Nesterov’s gradient method) and analyzed the non-ergodic convergence rate for the proposed method.

In this paper, we propose a new IAL framework (see Algorithm 1) for solving problem (1.1), where the AL subproblem is approximately solved until a point  $x^{k+1}$  is found such that

$$\max_{x \in \mathbb{R}^n} \left\{ \left\langle \nabla \hat{f}_\beta(x^{k+1}; \lambda^k), x^{k+1} - x \right\rangle + g(x^{k+1}) - g(x) \right\} \leq \eta_k. \quad (1.7)$$

Here  $\nabla \hat{f}_\beta(x; \lambda)$  is the gradient of  $\hat{f}_\beta(x; \lambda)$  with respect to  $x$ . The termination condition (1.7) in our proposed IAL framework is weaker and (potentially) easier to check<sup>2</sup> than (1.6) in most of the existing IAL frameworks/methods. We establish the global convergence of the proposed IAL framework under the assumption that the sequence  $\{\eta_k\}$  in (1.7) is summable; see Theorem 2.1. Moreover, we show, in Theorems 2.2 and 2.3, the non-ergodic convergence rate (under weaker conditions than that in [10]) for the proposed IAL framework, which reveals how the error in solving the AL subproblem affects the convergence rate.

## 2 The IAL framework and non-ergodic convergence rate analysis

In this section, we present the IAL framework for solving problem (1.1) and analyze its global convergence and non-ergodic convergence rate.

### 2.1 The IAL framework

The proposed IAL framework is given in Algorithm 1. At the  $k$ -th iteration, the IAL framework first solves AL subproblem (2.1) with fixed dual variable  $\lambda^k$  in an inexact manner until a point  $x^{k+1}$  satisfying (1.7) is found; then updates the dual variable by performing an inexact gradient ascent step (2.2).

<sup>2</sup> To check whether  $x^{k+1}$  satisfies (1.7) or not, we only need to solve the convex optimization problem on the left-hand side of (1.7), which can be solved exactly or to a high precision in time (essentially) linear in the size of the input for many  $g(x)$ ; see [7]. In contrast, it is generally hard to check whether  $x^{k+1}$  satisfies (1.6) or not (because  $x(\lambda^k)$  is unknown).

---

**Algorithm 1:** The IAL framework for problem (1.1)

---

```

1 Initialize  $x^1 \in \mathcal{X}$ ,  $\lambda^1 \in \mathbb{R}^m$ , and the nonnegative sequence  $\{\eta_k\}$ .
2 for  $k \geq 1$ : do
3   Find an approximate solution  $x^{k+1}$  of the AL subproblem
      
$$\min_{x \in \mathbb{R}^n} \left\{ \mathcal{L}_\beta(x; \lambda^k) := \hat{f}_\beta(x; \lambda^k) + g(x) \right\} \quad (2.1)$$

      such that (1.7) is satisfied;
4   Update the dual variable via
      
$$\lambda^{k+1} = \lambda^k + \beta (Ax^{k+1} - b). \quad (2.2)$$


```

---

Two remarks on the proposed IAL framework are in order.

First, AL subproblem (2.1) can be efficiently solved in an inexact manner by various (first-order) methods such as Nesterov's gradient methods [15–17] and the Frank-Wolfe (a.k.a. conditional gradient) methods [5, 8, 9, 18]. Specifically, Nesterov's gradient methods can find the point  $x^{k+1}$  satisfying (1.6) within  $\mathcal{O}\left(\frac{1}{\eta_k}\right)$  iterations. Since (1.6) implies (1.7) [14, Lemma 2.3], it follows that Nesterov's gradient methods can find the point  $x^{k+1}$  satisfying (1.7) within  $\mathcal{O}\left(\frac{1}{\eta_k}\right)$  iterations. The Frank-Wolfe methods can also find the point  $x^{k+1}$  satisfying (1.7) within  $\mathcal{O}\left(\frac{1}{\eta_k}\right)$  iterations. Compared to Nesterov's gradient methods, the Frank-Wolfe methods converge slower in practice, while the computational cost per iteration in the Frank-Wolfe methods is generally cheaper.

Second, the smaller the tolerance  $\eta_k$  is, the more computational cost is needed in Algorithm 1 to find the point  $x^{k+1}$  satisfying (1.7). On the other hand, the larger the tolerance  $\eta_k$  is, the larger the approximation error between the approximate gradient  $Ax^{k+1} - b$  and the true gradient  $\nabla d(\lambda^k)$  (see Lemma 2.6 further ahead) which might lead to slow convergence or even divergence of the proposed Algorithm 1. Therefore, the choice of  $\{\eta_k\}$  is important in balancing the computational cost (of finding the point  $x^{k+1}$  satisfying (1.7)) and the global convergence and convergence rate (of the framework). We will discuss the possible choices of  $\{\eta_k\}$  in the next subsection.

## 2.2 Global convergence and non-ergodic convergence rate

In this subsection, we present global convergence and non-ergodic convergence rate results of our IAL framework (Algorithm 1), which are regardless of the methods used to find the point  $x^{k+1}$  satisfying (1.7). More specifically, Theorem 2.1 shows the global convergence of the IAL framework under the assumption that the nonnegative sequence  $\{\eta_k\}$  is summable; and Theorems 2.2 and 2.3 show the non-ergodic convergence rate of the IAL framework.

**Theorem 2.1** *Let  $\{x^k\}$  and  $\{\lambda^k\}$  be generated by Algorithm 1. Suppose the non-negative sequence  $\{\eta_k\}$  satisfies*

$$\sum_{k=1}^{+\infty} \eta_k < +\infty. \quad (2.3)$$

*Then,*

$$\delta_k := d(\lambda^*) - d(\lambda^k) \rightarrow 0 \text{ and } \|Ax^{k+1} - b\| \rightarrow 0,$$

*where  $\lambda^*$  is an optimal solution to problem (1.3) and  $d(\lambda)$  is defined in (1.4).*

Theorem 2.1 shows the global convergence of Algorithm 1 under conditions (1.7) and (2.3). Classical conditions in [20] that guarantee the global convergence of the IAL framework are (1.6) and (2.3). Since (1.6) implies (1.7) by [14, Lemma 2.3], our conditions (1.7) and (2.3) are weaker than conditions (1.6) and (2.3) in [20].

Before presenting the non-ergodic convergence rate of Algorithm 1, we first define some notation. Let

$$\theta := \frac{\beta}{4B^2} \text{ and } B := \sqrt{\|\lambda^1 - \lambda^*\|^2 + 2\beta \sum_{k=1}^{+\infty} \eta_k} < +\infty. \quad (2.4)$$

Since  $\delta_k \rightarrow 0$  (cf. Theorem 2.1) and  $\eta_k \rightarrow 0$ , there exists  $k_0 \geq 4$  such that

$$\max_{k \geq k_0} \{\delta_k\} \leq \frac{1}{2\theta} \text{ and } \max_{k \geq k_0} \{\eta_k\} \leq \frac{1}{24\theta}. \quad (2.5)$$

Also, we define

$$\tau_1 := \frac{k_0}{4\theta} \text{ and } \tau_2 := \frac{1}{4\theta\sqrt{\eta_{k_0}}}. \quad (2.6)$$

It is easy to verify

$$\tau_1 \geq \frac{1}{\theta} \text{ and } \tau_2 \geq \sqrt{\frac{3}{2\theta}}. \quad (2.7)$$

**Theorem 2.2** *Let  $\{\lambda^k\}$  be generated by Algorithm 1. Suppose that the positive sequence  $\{\eta_k\}$  is nonincreasing and satisfies (2.3) and*

$$\sqrt{\frac{\eta_{k+1}}{\eta_k}} \geq \frac{k-2}{k}, \quad k = k_0, k_0 + 1, \dots, \quad (2.8)$$

*where  $k_0$  satisfies (2.5). Then,*

$$\delta_k \leq \frac{\tau_1}{k} + \tau_2\sqrt{\eta_k}, \quad k = k_0, k_0 + 1, \dots, \quad (2.9)$$

*where  $\tau_1$  and  $\tau_2$  are defined in (2.6).*

As shown in (2.9), the rate that  $\{\delta_k\}$  converges to zero depends on two terms, i.e.,  $\frac{\tau_1}{k}$  and  $\tau_2\sqrt{\eta_k}$ , and the rate is determined by the slower one of them: if  $k\sqrt{\eta_k} \rightarrow 0$ , then  $\delta_k \rightarrow 0$  with a rate of  $\mathcal{O}(1/k)$ ; otherwise,  $\delta_k \rightarrow 0$  with a rate of  $\mathcal{O}(\sqrt{\eta_k})$ . In particular, if  $\eta_k = 0$  for all  $k \geq 1$ , then Algorithm 1 reduces to the exact dual gradient ascent method, and achieves the  $\mathcal{O}(1/k)$  convergence rate.

These facts indicate that the sequence  $\{\eta_k\}$  in Algorithm 1 should not be chosen such that  $\{\sqrt{\eta_k}\}$  converges faster than  $\{1/k\}$  to zero. This is because that such a choice would increase the computational cost of solving the AL subproblem, but theoretically cannot improve the convergence rate of  $\{\delta_k\}$ , which is  $\mathcal{O}(1/k)$  in this case. One possible choice of the sequence  $\{\eta_k\}$  is

$$\eta_k = \frac{\sigma}{k^{2\alpha}}, \quad k = 1, 2, \dots \quad (2.10)$$

with some constant  $\sigma > 0$  and  $\alpha \in (\frac{1}{2}, 1]$ . It is easy to check that (2.10) satisfies all conditions required in Theorem 2.2.

Theorem 2.3 gives the non-ergodic convergence rate of Algorithm 1 when  $\eta_k$  is chosen as in (2.10).

**Theorem 2.3** *Let  $\{x^k\}$  and  $\{\lambda^k\}$  be generated by Algorithm 1. Suppose that the positive sequence  $\{\eta_k\}$  is chosen as in (2.10). Then,*

$$\delta_k \leq \frac{C}{k^\alpha}, \quad k = 1, 2, \dots, \quad (2.11)$$

where

$$C = 4\sqrt{\frac{3(\frac{3}{2}\theta\sigma + 1)\sigma}{\theta}} + \frac{4}{3}\max\left\{\delta_1, \frac{4}{\theta}\right\}, \quad (2.12)$$

and  $\theta$  is given in (2.4);

$$\|Ax^{k+1} - b\|^2 \leq \psi_k := \frac{2}{\beta} \left( \frac{C + \sqrt{\sigma}}{k^\alpha} \right), \quad k = 1, 2, \dots; \quad (2.13)$$

and

$$-\|\lambda^*\| \sqrt{\psi_k} - \frac{\beta}{2}\psi_k \leq F(x^{k+1}) - F(x^*) \leq (\|\lambda^*\| + B) \sqrt{\psi_k} + \eta_k, \quad k = 1, 2, \dots; \quad (2.14)$$

where  $\lambda^*$  is an optimal solution to problem (1.3) and  $B$  is given in (2.4).

As a direct consequence of Theorem 2.3, we obtain the following result.

**Corollary 2.4** *Let  $\{x^k\}$  and  $\{\lambda^k\}$  be generated by Algorithm 1 with  $\eta_k = \frac{\sigma}{k^2}$ . Then,*

$$\delta_k = \mathcal{O}\left(\frac{1}{k}\right), \quad \|Ax^{k+1} - b\|^2 = \mathcal{O}\left(\frac{1}{k}\right), \quad \text{and} \quad |F(x^{k+1}) - F(x^*)| = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right).$$

## 2.3 Proof of Theorems 2.1, 2.2, and 2.3

In this subsection, we prove Theorems 2.1, 2.2, and 2.3. For ease of presentation, we define

$$x(\lambda^k) := \arg \min_{x \in \mathbb{R}^n} \mathcal{L}_\beta(x; \lambda^k), \quad k = 1, 2, \dots, \quad (2.15)$$

$$\begin{aligned} \nabla d(\lambda^k) &:= Ax(\lambda^k) - b, \quad k = 1, 2, \dots, \\ \bar{d}(\lambda^k) &:= \mathcal{L}_\beta(x^{k+1}; \lambda^k), \quad k = 1, 2, \dots, \end{aligned} \quad (2.16)$$

$$\nabla \bar{d}(\lambda^k) := Ax^{k+1} - b, \quad k = 1, 2, \dots, \quad (2.17)$$

where  $x^{k+1}$  is generated by Algorithm 1 and satisfies (1.7).

We first prove the following two lemmas, which have been proved for smooth function  $F(x)$  in [3, 10, 14]. We now extend them to composite nonsmooth function  $F(x)$ . Lemma 2.5 shows that  $d(\lambda^{k+1})$  can be bounded from both above and below and Lemma 2.6 shows that  $\|\nabla d(\lambda^k) - \nabla \bar{d}(\lambda^k)\|$  is bounded by  $\sqrt{\eta_k/\beta}$ .

**Lemma 2.5** *The following two inequalities hold:*

$$d(\lambda) \leq \bar{d}(\mu) + \langle \nabla \bar{d}(\mu), \lambda - \mu \rangle, \quad \forall \lambda, \mu, \quad (2.18)$$

and

$$d(\lambda^{k+1}) \geq \bar{d}(\lambda^k) + \frac{\beta}{2} \|\nabla \bar{d}(\lambda^k)\|^2 - \eta_k. \quad (2.19)$$

*Proof* We first show (2.18). By the definitions of  $d(\lambda)$  and  $\mathcal{L}_\beta(x; \lambda)$ , we have

$$d(\lambda) = \min_{x \in \mathbb{R}^n} \{\mathcal{L}_\beta(x; \lambda)\} \leq \mathcal{L}_\beta(x_\mu; \lambda) = \mathcal{L}_\beta(x_\mu; \mu) + \langle \lambda - \mu, Ax_\mu - b \rangle,$$

where  $x_\mu$  satisfies  $\bar{d}(\mu) = \mathcal{L}_\beta(x_\mu; \mu)$ . This, together with the definition of  $\nabla \bar{d}(\mu)$  (cf. (2.16)), yields (2.18).

We now prove (2.19). By the convexity of  $f(x)$ , the definition of  $\nabla \bar{d}(\lambda^k)$ , and (2.2), we get

$$\begin{aligned} \mathcal{L}_\beta(x; \lambda^{k+1}) &\geq f(x^{k+1}) + \langle \nabla f(x^{k+1}), x - x^{k+1} \rangle + g(x) + \langle \lambda^{k+1}, Ax - b \rangle + \frac{\beta}{2} \|Ax - b\|^2 \\ &= \mathcal{L}_\beta(x^{k+1}; \lambda^k) + \beta \left( \langle \nabla \bar{d}(\lambda^k), Ax - b \rangle + \frac{1}{2} \|(Ax - b) - \nabla \bar{d}(\lambda^k)\|^2 \right) \\ &\quad + \langle \nabla \hat{f}_\beta(x^{k+1}; \lambda^k), x - x^{k+1} \rangle + g(x) - g(x^{k+1}). \end{aligned}$$

Taking the minimum over  $x \in \mathbb{R}^n$  on both sides of the above inequality, we have

$$\begin{aligned} d(\lambda^{k+1}) &\geq \mathcal{L}_\beta(x^{k+1}; \lambda^k) + \beta \min_{x \in \mathbb{R}^n} \left\{ \langle \nabla \bar{d}(\lambda^k), Ax - b \rangle + \frac{1}{2} \|(Ax - b) - \nabla \bar{d}(\lambda^k)\|^2 \right\} \\ &\quad + \min_{x \in \mathbb{R}^n} \left\{ \langle \nabla \hat{f}_\beta(x^{k+1}; \lambda^k), x - x^{k+1} \rangle + g(x) - g(x^{k+1}) \right\}. \end{aligned}$$

By using the definition of  $\bar{d}(\lambda)$  and (1.7), we immediately get the desired result (2.19).

**Lemma 2.6** *The following inequality holds:*

$$\left\| \nabla d(\lambda^k) - \nabla \bar{d}(\lambda^k) \right\|^2 \leq \frac{\eta_k}{\beta}. \quad (2.20)$$

*Proof* It follows from the optimality of  $x(\lambda^k)$  (cf. (2.15)) that

$$\left\langle \nabla \hat{f}_\beta(x(\lambda^k); \lambda^k), x(\lambda^k) - x^{k+1} \right\rangle + g(x(\lambda^k)) - g(x^{k+1}) \leq 0.$$

By setting  $x = x(\lambda^k)$  in (1.7), we get

$$\left\langle \nabla \hat{f}_\beta(x^{k+1}; \lambda^k), x^{k+1} - x(\lambda^k) \right\rangle - g(x(\lambda^k)) + g(x^{k+1}) \leq \eta_k.$$

Adding the above two inequalities yields

$$\begin{aligned} \eta_k &\geq \left\langle \nabla \hat{f}_\beta(x^{k+1}; \lambda^k) - \nabla \hat{f}_\beta(x(\lambda^k); \lambda^k), x^{k+1} - x(\lambda^k) \right\rangle \\ &= \left\langle \nabla f(x^{k+1}) + \beta A^T(Ax^{k+1} - b) - \nabla f(x(\lambda^k)) - \beta A^T(Ax(\lambda^k) - b), x^{k+1} - x(\lambda^k) \right\rangle \\ &\geq \left\langle \beta A^T(Ax^{k+1} - b) - \beta A^T(Ax(\lambda^k) - b), x^{k+1} - x(\lambda^k) \right\rangle \\ &= \beta \left\| \nabla d(\lambda^k) - \nabla \bar{d}(\lambda^k) \right\|^2, \end{aligned}$$

where the second inequality is due to the convexity of  $f(x)$  and the second equality is due to the definitions of  $\nabla d(\lambda^k)$  and  $\nabla \bar{d}(\lambda^k)$ .  $\square$

Lemma 2.7 shows that the sequence  $\{\lambda^k\}$  generated by Algorithm 1 is bounded.

**Lemma 2.7** *Let  $\{\lambda^k\}$  be generated by Algorithm 1. Suppose the sequence  $\{\eta_k\}$  satisfies (2.3), then*

$$\left\| \lambda^k - \lambda^* \right\| \leq B, \quad k = 1, 2, \dots, \quad (2.21)$$

where  $B$  is given in (2.4).

*Proof* We have

$$\begin{aligned} \left\| \lambda^{k+1} - \lambda^* \right\|^2 &= \left\| \lambda^k - \lambda^* + \beta \nabla \bar{d}(\lambda^k) \right\|^2 \\ &= \left\| \lambda^k - \lambda^* \right\|^2 + 2\beta \left\langle \nabla \bar{d}(\lambda^k), \lambda^k - \lambda^* \right\rangle + \beta^2 \left\| \nabla \bar{d}(\lambda^k) \right\|^2 \\ &\leq \left\| \lambda^k - \lambda^* \right\|^2 + 2\beta \left( \bar{d}(\lambda^k) - d(\lambda^*) \right) + \beta^2 \left\| \nabla \bar{d}(\lambda^k) \right\|^2 \\ &= \left\| \lambda^k - \lambda^* \right\|^2 + 2\beta \left( d(\lambda^{k+1}) - d(\lambda^*) \right) + 2\beta \left( \bar{d}(\lambda^k) - d(\lambda^{k+1}) \right) + \beta^2 \left\| \nabla \bar{d}(\lambda^k) \right\|^2 \\ &\leq \left\| \lambda^k - \lambda^* \right\|^2 + 2\beta \left( d(\lambda^{k+1}) - d(\lambda^*) \right) + 2\beta \eta_k \\ &\leq \left\| \lambda^k - \lambda^* \right\|^2 + 2\beta \eta_k, \end{aligned}$$

where the first inequality is due to (2.18) (with  $\lambda$  and  $\mu$  replaced by  $\lambda^*$  and  $\lambda^k$  respectively), the second inequality is due to (2.19), and the last inequality is due



to the fact that  $d(\lambda^{k+1}) \leq d(\lambda^*)$  for all  $k \geq 1$ . Summing the above inequality, we obtain

$$\|\lambda^k - \lambda^*\|^2 \leq \|\lambda^1 - \lambda^*\|^2 + 2\beta \sum_{i=1}^{k-1} \eta_i, \quad k = 1, 2, \dots,$$

which, together with (2.4), completes the proof.  $\square$

**Proof of Theorem 2.1.** To prove Theorem 2.1, it suffices to show

$$\sum_k \delta_k^2 < +\infty, \quad (2.22)$$

and

$$\sum_k \|Ax^{k+1} - b\|^2 < +\infty. \quad (2.23)$$

By (2.19) and the definition of  $\bar{d}(\lambda^k)$ , we obtain

$$d(\lambda^{k+1}) \geq d(\lambda^k) + \frac{\beta}{2} \|\nabla \bar{d}(\lambda^k)\|^2 - \eta_k, \quad (2.24)$$

which, together with (2.20) and the inequality  $a^2 \geq b^2/2 - (a-b)^2$ , implies

$$d(\lambda^{k+1}) \geq d(\lambda^k) + \frac{\beta}{4} \|\nabla d(\lambda^k)\|^2 - \frac{3}{2} \eta_k. \quad (2.25)$$

Moreover, it follows from (2.21) and the concavity of  $d(\lambda)$  that

$$d(\lambda^*) - d(\lambda^k) \leq \langle \nabla d(\lambda^k), \lambda^* - \lambda^k \rangle \leq \|\lambda^k - \lambda^*\| \|\nabla d(\lambda^k)\| \leq B \|\nabla d(\lambda^k)\|.$$

Combining the above and (2.25), we immediately obtain

$$\delta_{k+1} \leq \delta_k - \theta \delta_k^2 + \frac{3}{2} \eta_k, \quad k = 1, 2, \dots, \quad (2.26)$$

which further implies

$$\delta_k \leq \delta_1 - \theta \sum_{i=1}^{k-1} \delta_i^2 + \frac{3}{2} \sum_{i=1}^{k-1} \eta_i, \quad k = 1, 2, \dots \quad (2.27)$$

From the definition of  $\delta_k$ , we know  $\delta_k \geq 0$  for all  $k \geq 1$ . From this, (2.3), and (2.27), we obtain (2.22).

Next, we prove (2.23). It follows from (2.17) and (2.24) that

$$\|Ax^{k+1} - b\|^2 \leq \frac{2}{\beta} \left( d(\lambda^{k+1}) - d(\lambda^k) + \eta_k \right), \quad k = 1, 2, \dots \quad (2.28)$$

Summing (2.28) from  $i = 1$  to  $k$  yields

$$\begin{aligned}
\sum_{i=1}^k \|Ax^{i+1} - b\|^2 &\leq \frac{2}{\beta} \left( d(\lambda^{k+1}) - d(\lambda^1) + \sum_{i=1}^k \eta_i \right) \\
&\leq \frac{2}{\beta} \left( d(\lambda^*) - d(\lambda^1) + \sum_{i=1}^k \eta_i \right) \\
&\leq \frac{2}{\beta} \left( \frac{1}{2\beta} \|\lambda^1 - \lambda^*\|^2 + \sum_{i=1}^k \eta_i \right) \\
&\leq \frac{B^2}{\beta^2},
\end{aligned}$$

where the last second inequality is due to the facts that  $\nabla d(\lambda^*) = 0$  and  $\nabla d(\lambda)$  is  $\frac{1}{\beta}$ -Lipschitz continuous [1] and the last inequality is due to (2.4). The proof of Theorem 2.1 is completed.  $\square$

**Proof of Theorem 2.2.** We prove the theorem by induction. From (2.5) and (2.6), we know

$$\delta_{k_0} \leq \frac{1}{2\theta} = \frac{\tau_1}{k_0} + \tau_2 \sqrt{\eta_{k_0}}.$$

Therefore, the inequality (2.9) holds for  $k = k_0$ . Next, we assume that (2.9) holds for some  $k \geq k_0$ , and we consider the case  $k + 1$ . We have

$$\begin{aligned}
\delta_{k+1} &\leq \delta_k - \theta \delta_k^2 + \frac{3}{2} \eta_k \\
&\leq \frac{\tau_1}{k} + \tau_2 \sqrt{\eta_k} - \theta \left( \frac{\tau_1}{k} + \tau_2 \sqrt{\eta_k} \right)^2 + \frac{3}{2} \eta_k \\
&= \frac{\tau_1}{k+1} \frac{(k+1)(k-\theta\tau_1)}{k^2} + \tau_2 \sqrt{\eta_{k+1}} \frac{\sqrt{\eta_k} \left(1 - \frac{2\theta\tau_1}{k}\right)}{\sqrt{\eta_{k+1}}} + \left( \frac{3}{2} - \theta\tau_2^2 \right) \eta_k \\
&\leq \frac{\tau_1}{k+1} + \tau_2 \sqrt{\eta_{k+1}},
\end{aligned}$$

where the first inequality is due to (2.26), the second inequality is due to the fact that  $\{\eta_k\}$  is nonincreasing, which further implies  $\frac{\tau_1}{k} + \tau_2 \sqrt{\eta_k} \leq \frac{\tau_1}{k_0} + \tau_2 \sqrt{\eta_{k_0}} = \frac{1}{2\theta}$  for all  $k \geq k_0$ , and the last inequality is due to (2.7) and (2.8). The proof of Theorem 2.2 is completed.  $\square$

To prove Theorem 2.3, we need the following lemma.

**Lemma 2.8** *Suppose the nonnegative sequence  $\{\delta_k\}$  satisfies*

$$\frac{E}{2} \delta_{k+1}^2 + \delta_{k+1} \leq \delta_k, \quad k = 1, 2, \dots, \quad (2.29)$$

where  $E > 0$  is a constant. Then, we have

$$\delta_k \leq \frac{\max\{\delta_1, \frac{4}{E}\}}{k}, \quad k = 1, 2, \dots \quad (2.30)$$

*Proof* Clearly, the inequality (2.30) is true for  $k = 1$ . Next, assuming (2.30) is true for some  $k \geq 1$ , we show it is also true for  $k + 1$ . In fact, we have

$$\begin{aligned} \delta_{k+1} &\leq \frac{-1 + \sqrt{1 + 2E\delta_k}}{E} \leq \frac{-1 + \sqrt{1 + 2E \frac{\max\{\delta_1, \frac{4}{E}\}}{k}}}{E} \\ &= \frac{2 \max\{\delta_1, \frac{4}{E}\}}{k + \sqrt{k^2 + 2E \max\{\delta_1, \frac{4}{E}\}} k} \\ &\leq \frac{\max\{\delta_1, \frac{4}{E}\}}{k + 1}, \end{aligned}$$

where the first inequality is due to the inequality (2.29), and the second inequality is due to the assumption that (2.30) holds for  $k$ .  $\square$

**Proof of Theorem 2.3.** We show (2.11), (2.13), and (2.14) separately.

We first show (2.11). Clearly, the inequality (2.11) holds for  $k = 1$ . Next, we assume that (2.11) holds for some  $k \geq 1$ , and show it is also true for  $k + 1$ . We use the contrapositive argument. Assume that (2.11) does not hold for  $k + 1$ , i.e.,

$$\delta_{k+1} > \frac{C}{(k+1)^\alpha}, \quad (2.31)$$

where  $C$  is given in (2.12). Let  $z^* = \frac{C}{4} - \frac{1}{\theta} > 0$ . If

$$\frac{\theta}{2} \left( \delta_{k+1} - \frac{z^*}{(k+1)^\alpha} \right)^2 + \left( \delta_{k+1} - \frac{z^*}{(k+1)^\alpha} \right) \leq \left( \delta_k - \frac{z^*}{k^\alpha} \right) \quad (2.32)$$

holds, then it follows from Lemma 2.8 that

$$\delta_{k+1} - \frac{z^*}{(k+1)^\alpha} \leq \frac{\max\{\delta_1, \frac{4}{\theta}\}}{k+1},$$

and

$$\delta_{k+1} \leq \frac{z^* + \max\{\delta_1, \frac{4}{\theta}\}}{(k+1)^\alpha} < \frac{C}{(k+1)^\alpha}. \quad (2.33)$$

Clearly, (2.33) contradicts (2.31), which implies that (2.11) is true. Next, we prove (2.32), which is equivalent to

$$P(z^*) := \frac{\theta}{2} \left( \delta_{k+1} - \frac{z^*}{(k+1)^\alpha} \right)^2 + \left( \delta_{k+1} - \frac{z^*}{(k+1)^\alpha} \right) - \left( \delta_k - \frac{z^*}{k^\alpha} \right) \leq 0.$$

We consider the following quadratic function  $Q(z)$  with respect to  $z$ :

$$Q(z) = \frac{\theta}{2} z^2 - \left( \frac{\theta C}{4} - 1 \right) z + \frac{3}{2} \left( \frac{3\theta\sigma}{2} + 1 \right) \sigma.$$

It can be verified that the minimizer of  $Q(z)$  is  $z^*$ . Since the discriminant of  $Q(z)$  is nonnegative, it follows that the minimum value

$$Q(z^*) \leq 0. \quad (2.34)$$

Moreover, for any  $k \geq 1$ , we have

$$(2.26) \implies \delta_{k+1} \leq \delta_k + \frac{3}{2}\eta_k \implies \frac{1}{2}\delta_{k+1}^2 \leq \delta_k^2 + \frac{9}{4}\eta_k^2 \implies -\delta_k^2 \leq -\frac{1}{2}\delta_{k+1}^2 + \frac{9}{4}\sigma\eta_k.$$

Combining the last inequality in the above with (2.26) yields

$$\frac{\theta}{2}\delta_{k+1}^2 + \delta_{k+1} \leq \delta_k + \frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\eta_k, \quad (2.35)$$

which further implies

$$\begin{aligned} P(z^*) &= \frac{\theta}{2}\delta_{k+1}^2 + \delta_{k+1} - \delta_k - \frac{\theta\delta_{k+1}z^*}{(k+1)^\alpha} + \frac{\theta(z^*)^2}{2(k+1)^{2\alpha}} - \frac{z^*}{(k+1)^\alpha} + \frac{z^*}{k^\alpha} \\ &\leq \frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\eta_k - \frac{\theta\delta_{k+1}z^*}{(k+1)^\alpha} + \frac{\theta(z^*)^2}{2(k+1)^{2\alpha}} - \frac{z^*}{(k+1)^\alpha} + \frac{z^*}{k^\alpha}. \end{aligned} \quad (2.36)$$

By  $\eta_k \leq \frac{\sigma}{k^{2\alpha}}$ , the assumption  $\delta_{k+1} > \frac{C}{(k+1)^\alpha}$  (cf. (2.31)), the facts  $(k+1)^{2\alpha} \leq 4k^{2\alpha}$  and  $(k+1)^\alpha - k^\alpha \leq 1$  for all  $k \geq 1$  and  $\alpha \in (0, 1]$ , we get

$$\begin{aligned} &\frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\eta_k - \frac{\theta\delta_{k+1}z^*}{(k+1)^\alpha} + \frac{\theta(z^*)^2}{2(k+1)^{2\alpha}} - \frac{z^*}{(k+1)^\alpha} + \frac{z^*}{k^\alpha} \\ &\leq \frac{\frac{3}{2}\left(\frac{3\theta\sigma}{2} + 1\right)\sigma}{k^{2\alpha}} - \frac{\theta C}{4k^{2\alpha}}z^* + \frac{\theta}{2k^{2\alpha}}(z^*)^2 + \frac{1}{k^{2\alpha}}z^* \\ &= \frac{Q(z^*)}{k^{2\alpha}}, \end{aligned}$$

which, together with (2.34) and (2.36), yields  $P(z^*) \leq 0$ .

We now show (2.13). From (2.28), we obtain

$$\begin{aligned} \|Ax^{k+1} - b\|^2 &\leq \frac{2}{\beta} \left( d(\lambda^{k+1}) - d(\lambda^k) + \eta_k \right) \leq \frac{2}{\beta} \left( d(\lambda^*) - d(\lambda^k) + \eta_k \right) \\ &= \frac{2}{\beta} (\delta_k + \eta_k), \end{aligned}$$

which, together with (2.10) and (2.11), yields (2.13).

Finally, we show (2.14). From the strong duality and the definition of  $\mathcal{L}_\beta(x; \lambda)$  (cf. (1.2)), we obtain

$$\begin{aligned} F(x^*) &\leq \mathcal{L}_\beta(x^{k+1}; \lambda^*) = F(x^{k+1}) + \left\langle \lambda^*, Ax^{k+1} - b \right\rangle + \frac{\beta}{2} \|Ax^{k+1} - b\|^2 \\ &\leq F(x^{k+1}) + \|\lambda^*\| \|Ax^{k+1} - b\| + \frac{\beta}{2} \|Ax^{k+1} - b\|^2. \end{aligned}$$

This, together with (2.13), implies

$$F(x^{k+1}) - F(x^*) \geq -\|\lambda^*\| \sqrt{\psi_k} - \frac{\beta}{2} \psi_k. \quad (2.37)$$

On the other hand, we have

$$\begin{aligned}
\mathcal{L}_\beta(x^{k+1}; \lambda^k) &\leq \hat{f}_\beta(x(\lambda^k); \lambda^k) + \left\langle \nabla \hat{f}_\beta(x^{k+1}; \lambda^k), x^{k+1} - x(\lambda^k) \right\rangle + g(x^{k+1}) \\
&= d(\lambda^k) + \left\langle \nabla \hat{f}_\beta(x^{k+1}; \lambda^k), x^{k+1} - x(\lambda^k) \right\rangle + g(x^{k+1}) - g(x(\lambda^k)) \\
&\leq d(\lambda^k) + \eta_k \\
&\leq F(x^*) + \eta_k,
\end{aligned}$$

where the first inequality is due to the convexity of  $\hat{f}_\beta(x; \lambda)$  with respect to  $x$ , the first equality is due to the definition of  $d(\lambda^k)$ , the second inequality is due to (1.7), and the last inequality is due to the fact  $d(\lambda^k) \leq F(x^*)$ . Recall the definition of  $\mathcal{L}_\beta(x; \lambda)$ , we get

$$F(x^{k+1}) + \left\langle \lambda^k, Ax^{k+1} - b \right\rangle + \frac{\beta}{2} \|Ax^{k+1} - b\|^2 \leq F(x^*) + \eta_k,$$

which, together with (2.21), immediately implies

$$F(x^{k+1}) - F(x^*) \leq (\|\lambda^*\| + B) \sqrt{\psi_k} + \eta_k. \quad (2.38)$$

Combining (2.37) and (2.38) yields (2.14). The proof of Theorem 2.3 is completed.  $\square$

## 2.4 Remarks

In this subsection, we make some remarks on the comparison of our proposed IAL framework (Algorithm 1) and one closely related IAL method [10] for solving the linearly constrained convex programming.

The method in [10] is designed for solving problem (1.1) with  $g(x) = \text{Ind}_\mathcal{X}(x)$ . They apply Nesterov's optimal first-order method to solve AL subproblem (2.1) until a point  $x^{k+1}$  satisfying (1.6) is found. Our IAL framework can be used to solve more general problem (1.1) (with a general composite function  $g(x)$ ). Our framework requires approximately solving subproblem (2.1) until a point  $x^{k+1}$  satisfying (1.7) (which is easier to check than (1.6)) is found.

The work [10] shows the same non-ergodic convergence rate results as ours in Corollary 2.4 but under a much stronger condition that the sequence  $\{\eta_k\}$  in (1.6) satisfies

$$\sum_{i=1}^k \eta_i^2 = \mathcal{O}\left(\frac{1}{k}\right).$$

To make it more clearly, consider the special case where we are interested in finding an exact solution of problem (1.1), which requires  $k \rightarrow +\infty$  in Corollary 2.4. In this case, the method in [10] needs solving each AL subproblem exactly (i.e.,  $\eta_i$  in (1.6) needs to be zero for all  $i = 1, 2, \dots$ ) while our IAL framework only needs solving each subproblem approximately (i.e.,  $\eta_i$  in (1.7) only needs to be in the order of  $\mathcal{O}(1/i^2)$  for  $i = 1, 2, \dots$ ).

**Acknowledgements** We would like to thank Guanghui Lan, Zhaosong Lu, Zaiwen Wen, and Wotao Yin for their insightful comments, which helped us in improving the results in this paper. We thank Xiangfeng Wang for the useful discussion on an earlier version of this paper.

## References

1. Bertsekas, D.P.: *Nonlinear Programming*. Athena Scientific, Belmont, MA (1999)
2. Bertsekas, D.P., Tsitsiklis, J.N.: Gradient convergence in gradient methods with errors. *SIAM J. Optim.* **10**(3), 627–642 (2000)
3. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. *Math. Program.* **146**(1-2), 37–75 (2014)
4. Eckstein, J., Silva, P.J.: A practical relative error criterion for augmented Lagrangians. *Math. Program.* **141**(1-2), 319–348 (2013)
5. Frank, M., Wolfe, P.: An algorithm for quadratic programming. *Naval Research Logistics Quarterly* **3**(1-2), 95–110 (1956)
6. Hestenes, M.R.: Multiplier and gradient methods. *J. Optim. Theory Appl.* **4**(5), 303–330 (1969)
7. Jaggi, M.: *Sparse convex optimization methods for machine learning*. Ph.D. thesis, ETH Zurich (2011)
8. Jaggi, M.: Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In: *Proceeding of ICML, Atlanta* (2013)
9. Lan, G.: Gradient sliding for composite optimization. *Math. Program.* (2015). DOI 10.1007/s10107-015-0955-5
10. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Math. Program.* (2015). DOI 10.1007/s10107-015-0861-x
11. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: A general approach. *Ann. Oper. Res.* **46**(1), 157–178 (1993)
12. Necoara, I., Patrascu, A.: Iteration complexity analysis of dual first order methods for conic convex programming. <http://arxiv.org/abs/1409.1462> (2015)
13. Necoara, I., Patrascu, A., Glineur, F.: Complexity certifications of first order inexact Lagrangian and penalty methods for conic convex programming. <http://arxiv.org/abs/1506.05320> (2015)
14. Nedelcu, V., Necoara, I., Tran-Dinh, Q.: Computational complexity of inexact gradient augmented Lagrangian methods: Application to constrained MPC. *SIAM J. Control Optim.* **52**(5), 3109–3134 (2014)
15. Nesterov, Y.: A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ . *Sov. Math. Dokl.* **27**(2), 372–376 (1983)
16. Nesterov, Y.: Smooth minimization of non-smooth functions. *Math. Program.* **103**(1), 127–152 (2005)
17. Nesterov, Y.: Gradient methods for minimizing composite functions. *Math. Program.* **140**(1), 125–161 (2013)
18. Nesterov, Y.: Complexity bounds for primal-dual methods minimizing the model of objective function. <http://www.uclouvain.be/en-452691.html> (2015)
19. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: R. Fletcher (ed.) *Optimization* pp. 283–298 (1969)
20. Rockafellar, R.T.: Augmented Lagrangian and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.* **1**(2), 97–116 (1976)
21. So, A.M.C.: Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. <http://arxiv.org/abs/1309.0113> (2013)
22. Yang, L., Sun, D., Toh, K.C.: SDPNAL+: A majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints. *Math. Prog. Comp.* (2015). DOI 10.1007/s12532-015-0082-6